Journal of Biomedical Informatics 42 (2009) 605–611

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Analyzing time-dependent microarray data using independent component analysis derived expression modes from human macrophages infected with *F. tularensis holartica*

D. Lutter [a,b,e,*], Th. Langmann [a,d], P. Ugocsai [a], C. Moehle [a], E. Seibold [c], W.D. Splettstoesser [c], P. Gruber [b], E.W. Lang [b], G. Schmitz [a]

[a] *Clinical Chemistry, University Clinic, 93053 Regensburg, Germany*
[b] *CIML Group, Institute of Biophysics, University of Regensburg, 93040 Regensburg, Germany*
[c] *Bundeswehr Institute of Microbiology, Neuherbergstr. 11, 80937 Munich, Germany*
[d] *Institute of Human Genetics, University Clinic, 93053 Regensburg, Germany*
[e] *Institute of Bioinformatics and Systems Biology, CMB, Helmholtz Zentrum Muenchen, Germany*

ARTICLE INFO

Article history:
Received 8 May 2008
Available online 23 January 2009

Keywords:
Independent component analysis
Microarray
Gene chip
Infection
Time course
Clustering
Francisella tularensis

ABSTRACT

The analysis of large-scale gene expression profiles is still a demanding and extensive task. Modern machine learning and data mining techniques developed in linear algebra, like Independent Component Analysis (ICA), become increasingly popular as appropriate tools for analyzing microarray data. We applied ICA to analyze kinetic gene expression profiles of human monocyte derived macrophages (MDM) from three different donors infected with *Francisella tularensis holartica* and compared them to more classical methods like hierarchical clustering. Results were compared using a pathway analysis tool, based on the Gene Ontology and the MeSH database. We could show that both methods lead to time-dependent gene regulatory patterns which fit well to known TNFα induced immune responses. In comparison, the nonexclusive attribute of ICA results in a more detailed view and a higher resolution in time dependent behavior of the immune response genes. Additionally, we identified NFκB as one of the main regulatory genes during response to *F. tularensis* infection.

1. Introduction

Environmental stimuli or the activity of the internal state of cells induce or repress genes via up- or down-regulation of corresponding expressed mRNAs. Gene expression is controlled by a combination of mechanisms including those involving networks of signaling molecules, transcription factors and their binding sites in the promotor regions of genes, as well as modifications of the chromatin structure and different types of post-transcriptional regulation. The expression of each gene thus relies on the specific processing of a number of regulatory inputs.

High-throughput genome-wide measurements of transcript levels have become available with the recent development of microarray technology [1]. Intelligent and efficient mathematical and computational analysis tools are needed to read and interpret the information content buried in these large data sets (for a recent review see [2,3]).

Traditionally two strategies exist to analyze such data sets. If prior knowledge about classification of the samples is available, a *supervised*, also called *knowledge-based*, analysis can identify gene expression patterns, called features, specific to a given class, which can be used to classify new samples. Without any hypothesis, *unsupervised*, i.e. data driven, approaches can discover novel biological mechanisms and reveal genetic regulatory networks in large data sets. Such unsupervised analysis methods for microarray data analysis can be divided into clustering approaches, model-based approaches and projection methods. Clustering approaches group genes by some measure of similarity. A fundamental assumption of such clustering approaches is that genes within a cluster are functionally related. In general, no attempt is made to model the underlying biology. A drawback of such classical methods is that clusters generally are disjunct but genes may be part of several biological processes. Model-based approaches try to explain the interactions among the biological entities with the help of hypothesized concepts. Parameters of the model can be trained from expression data sets [12]. With complex models not enough data may be available to properly estimate the parameters, hence overfitting may result. Projective subspace methods try to expand

* Corresponding author. Address: Institute of Bioinformatics and Systems Biology, CMB, Helmholtz Zentrum Muenchen, Germany.
E-mail address: dominik.lutter@helmholtz-muenchen.de (D. Lutter).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.01.002

the data in a basis with desired properties. Projective subspace methods commonly used are principal component analysis (PCA), independent component analysis (ICA) or non-negative matrix factorization (NMF). Note that often PCA is a necessary preprocessing step for ICA algorithms. Here we focus on the well-known stochastic FastICA algorithm to analyze our time-dependent gene expression profiles (GEPs).

ICA decomposes the GEPs into statistically independent *gene expression modes* (GEM), the so-called independent components (ICs) [5]. The algorithm FastICA assumes a linear superposition of these unknown GEMs, also called source signals, forming the observed GEPs measured with microarray gene chips. Each retrieved GEM is considered to reflect a basic building block of a putative regulatory process, which can be characterized by the functional annotations of the genes that are predominant within the component. Each GEM thus defines corresponding groups of induced and/or repressed genes. Genes can be visualized by projecting them to particular expression modes which help to highlight particular biological functions, to reduce noise, and to compress the data in a biologically meaningful way.

In this work microarray data of human macrophages, deduced from human monocytes by M-CSF triggered differentiation and infected with a *Francisella tularensis holartica* strain called LVS (live vaccine strain), were analyzed. Our aim was to determine the global gene expression profile of human macrophages from three different donors infected *in vitro* with *F. tularensis* LVS. Expression profiles were followed over a period of 72 h, resulting in a series of ten experiments. To monitor assay and hybridization performance, a set of quality parameters (poly-A controls, hybridization controls, percent present, background and noise values, scaling factor) were assessed. None of them exceeded the given ranges, indicating that our data is of high quality. An analysis of these experiments using the FastICA algorithm [7] is reported in this work.

2. Methods

2.1. Sample preparation and expression level calculation

Human monocytes were obtained from three healthy donors by diagnostic leukapheresis and counterflow elutriation as described previously [10] under full GLP (good laboratory practice) conditions. The cells were cultured on plastic petri dishes in macrophage SFM medium (Gibco BRL, Karlsruhe) and allowed to differentiate for 5 days in the presence of 50 ng/ml recombinant human M-CSF (R&D Systems, Wiesbaden, Germany) to macrophages. Finally, the cells were infected with *F. tularensis* LVS. Three independent *F. tularensis* LVS infection experiments were chosen for further analysis. The infection rates and the percentage of living cells were comparable in all three experiments.

Total RNA was extracted from cultured cells according to the manufacturer's instructions using the RNeasy Protect Midi Kit (Qiagen, Hilden, Germany). Purity and integrity of the RNA was assessed on the Agilent 2100 bioanalyzer with the RNA 6000 Nano LabChip® reagent set (Agilent Technologies, USA). The RNA was quantified spectrophotometrically and then stored at −80 °C. At each timepoint enough total RNA could be isolated for DNA-microarray analysis and subsequent realtime RT-PCR verification experiments. The quality assessment of RNA samples is a major point in DNA-microarray analysis. All RNAs were of superior quality without any signs of mRNA degradation. The RNA integrity number (RIN) was close to the optimum (10) in all experiments.

Gene expression levels were measured using Affymetrix GeneChip® HGU133 Plus 2.0 Arrays. Array comparison analysis was carried out by calculating expression levels and fold changes using Affymetrix GeneChip Operating Software (GCOS). Expression values after 0.5, 1, 2, 3, 6, 9 and 12 h of incubation with 100 MOI (multiplicity of infection) *F. tularensis* LVS were compared to the 1 h control incubation. Furthermore, infected and control probes were compared after incubation at 24, 48 and 72 h.

2.2. Model assumptions

The transcription level of all genes in a cell is the result of the action of several regulatory processes which in parallel control the response of a cell to external stimuli. Matrix decomposition techniques set out to factorize a set of observed GEPs into components according to some specified constraints to assure unique decompositions. Such constraints then lead to either *statistically uncorrelated* (PCA) or even *statistically independent* (ICA) components. The latter may often be identified as regulatory processes governed by signaling pathways which are only weakly coupled to each other and can be considered as acting independently of each other to a first approximation. Each such process can then be represented by a vector of expression levels of up- or down-regulated genes, the *gene expression modes* (GEMs). Under each experimental condition, the different regulatory processes then linearly superimpose the expression levels of each gene according to the different GEMs to result in the observed GEPs measured by a microarray sample. The justification of such simplifying assumptions comes from the "biological meaning" of the resulting expression modes extracted by such matrix decomposition techniques. If such GEMs can clearly be identified with known signaling pathways within a cell for the problem at hand, the model decomposition is justified. Otherwise non-linear decompositions might need to be considered. For such matrix factorization algorithms to be applied, centered data, i.e. $\langle \mathbf{x} \rangle = 0$, will be assumed for simplicity. This can always be achieved by subtracting a time averaged expression level from each data point.

2.3. ICA model

Given the state of a cell at the time of experiment is governed by M regulatory processes $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_M)^T$ which are considered reasonably independent of each other and operate in parallel, and where each of them is represented by a row vector of K gene expression levels, i.e. $\mathbf{s}_m = (s_{m1}, \ldots, s_{mK})$, then $\mathbf{S}$ forms a $M \times K$ matrix whose rows consist of statistically independent GEMs. Each such mode forms a component expression pattern or component signature, in which the contribution of each gene to the envisaged independent regulatory processes is reflected via its expression level. Within a microarray experiment, the level of expression of all genes $\mathbf{x}_n = (x_{n1}, \ldots, x_{nK})$ is measured under N different experimental conditions, resulting in a microarray expression matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$, where the rows form the GEPs $\mathbf{x}_n$. Hence, a microarray data matrix $\mathbf{X}$ can be formed with N rows, representing GEPs, and K columns, representing the expression levels of a gene across all experimental conditions. Assuming that different experimental conditions cause different expression levels of each gene within the independent regulatory processes, each observed GEP, i.e. each row of $\mathbf{X}$, results as a weighted superposition of the independent GEMs, represented by the rows of $\mathbf{S}$. In matrix notation this model then reads

$$\mathbf{X} = \mathbf{APDS}, \tag{1}$$

where $\mathbf{A}$ represents the $N \times M$ matrix of mixing coefficients and here we set $N = M$. The *under-determined* or *over-determined* cases with $N \neq M$ is more difficult and will not be considered here. The N columns of $\mathbf{A}$ may be considered to form a new representation with basis vectors $\mathbf{a}_m = (a_{m1}, \ldots, a_{mN})$, also called feature profiles (FP), where each a_{mn} defines the weight with which the nth GEM

contributes to the mth observed GEP. In addition, the matrices $\mathbf{P}$ and $\mathbf{D}$ account for trivial permutation and scaling indeterminacies.

By approximating the negentropy as a measure of statistical independence, the FastICA [7] algorithm computes a de-mixing matrix $\mathbf{W}$ such that

$$\mathbf{Y} = \mathbf{WX}, \tag{2}$$

where $\mathbf{Y}$ represents a matrix of transformed variables $\mathbf{y}_1, \ldots, \mathbf{y}_N$, which correspond to the extracted independent components or GEMs subject to scaling ($\mathbf{D}$) and permutation ($\mathbf{P}$) indeterminacies [8]. They are extracted from the data by the algorithm as statistically independent as possible, and represent close approximations of the unknown expression signatures of the hypothetical underlying regulatory processes represented by $\mathbf{s}_1, \ldots, \mathbf{s}_N$.

2.4. Stability analysis

The number of GEMs extracted by the FastICA algorithm corresponds to the number of experiments, i.e. the number of different microarray data sets available. As the number of underlying independent regulatory processes contributing to any observed set of expression signatures is generally unknown, the GEMs extracted, due to the independence constraint enforced by the data matrix decomposition, may, at least to some extent, still represent superpositions of such underlying regulatory processes being searched for. This fact results in fluctuations in the estimated GEM upon repeated decomposition of the given data matrix. Unfortunately, these fluctuations also sometimes confounds the immediate and straightforward biological interpretation of such modes. Despite this it is the hope of every matrix decomposition analysis that the resulting GEMs provide for a more intuitive and insightful interpretation of the observed states of the cell under the experimental conditions and environmental stimuli to which it was exposed.

Because FastICA belongs to the class of stochastic matrix decomposition algorithms, the robustness of its results needs to be assured. To test the robustness of the resulting GEMs, we performed a bootstrap analysis. To do so, we randomly generated 50 sub-samples with a sample size 25% smaller than the original data set. As a consequence, repeating the analysis $L = 50$ times might render some or all of the extracted components to differ slightly in the various repeats. We then estimated the robustness of these repeatedly extracted GEMs.

We combined the rows $\mathbf{w}_n^l$ to a set $\mathscr{W}$ of row vectors, where l represents a particular ICA run and n is the nth row of the de-mixing matrix $\mathbf{W}^l$. Because $\mathbf{W} = \mathbf{A}^{-1}$ each row vector $\mathbf{w}_n$ contains the weights with which each observed GEP is combined to an extracted GEM. Using a projective *k-means* clustering [17] the resulting row vectors are then clustered into N clusters according to the following metric representing our distance or similarity measure:

$$d(\mathbf{w}, v) := \sqrt{1 - \left(\frac{\mathbf{w}^T v}{\sqrt{\|\mathbf{w}\|\|v\|}}\right)^2} \quad \mathbf{w}, v \in \mathscr{W}. \tag{3}$$

Now we use the centers of gravity of each cluster as code book vectors $\mathbf{c}_n, n = 1, \ldots, N$ for our stability analysis. The result of the clustering can be described by the sets $\mathscr{W}_n = \{\mathbf{w} \in \mathscr{W} | s(\mathbf{w}) = \mathbf{c}_n\}$ with $s(\mathbf{w}) = \arg\min_n d(\mathbf{w}, \mathbf{c}_n)$.

We evaluated the quality of each cluster $\mathscr{W}_n$ by calculating the 1^{st} and 2^{nd} moment of the distance distribution within each cluster, i.e. the empirical mean and standard deviation of all distances between the code book vector $\mathbf{c}_n$ of cluster n and the data vectors within the cluster using the distance measure d as defined above. In particular, $\text{mean}_n = \text{mean}(\{d(\mathbf{w}, \mathbf{c}_n) | \mathbf{w} \in \mathscr{W}_n\})$ and $\text{var}_n = \text{var}(\{d(\mathbf{w}, \mathbf{c}_n) | \mathbf{w} \in \mathscr{W}_n\})$ (Fig. 1). As a null model we randomly sampled N clusters from $\mathscr{W}$ with size L. For each sampled cluster we calculated the mean and standard deviation of all distances between the sampled vectors and the respective projective centroid.

2.5. Grouping genes

Each estimated GEM contains the gene expression levels of all genes within any given microarray experiment, i.e. every experimental condition chosen. Assuming that the genes involved in a hypothetical regulatory process represented by the GEM show relatively high expression within this GEM, then those genes are of utmost interest which correspond to the most or the least expressed. Only genes whose expression level exceeded the mean expression level plus five times the standard deviation of the considered GEM were retained for further analysis. These genes have been grouped together into gene groups of size between 35 and 94 genes, containing the most strongly expressed or suppressed genes. Remember that one gene may be involved in more than one regulatory process, i.e. its expression level may be high or low in several gene expression modes.

2.6. Biological relevance

Further information about the biological relevance of the genes and their regulation mechanisms can be gathered from public databases such as *Gene Ontology* (GO) (available at http://www.geneontology.org/). The biological information available within GO can be further explored using software tools like *Onto-Express* [3,9] (available at http://vortex.cs.wayne.edu/Projects.html) or *Genomatix BiblioSphere* (see http://www.genomatix.de/).

BiblioSphere provides further biological information by structuring input data into biological pathways, i.e. networks of interacting genes thereby delivering systems biology knowledge to organize genes within groups into functional networks. The interaction network is a data-mining solution in which relationships from the literature databases, genome-wide promoter analysis and verified gene interactions are combined. Results can be classified by tissue, Gene Ontology and MeSH (see http://www.nlm.nih.gov/mesh/).

Statistical rating by Z-scores indicate over- and under-representation of genes in the certain biological categories which are organized into hierarchies. For each term in the hierarchy, a statistical analysis is performed based on the number of observed and expected annotations. With each associated GO or MeSH term a Z-score is provided measuring the relevance of the functional term within the context of the group of genes under consideration. Z-scores are given by $Z\text{-score} = (n - \hat{n})/\sigma_n$, where n is the number of observed genes meeting any given criterion, $\hat{n}$ is the corresponding expected number and the standard deviation σ_n measures the fluctuations of n around the mean. The Z-score of this term helps to estimate whether a certain annotation, or group of annotations, is over- or under-represented in the tested set. Such score helps to determine whether the accumulation of annotations in a certain branch of the hierarchy is meaningful.

3. Results

3.1. Pathways biostatistics

For a knowledge-based pathway analysis, all expressed genes from the three LVS infection experiments were mapped to 78 manually annotated biomedical pathways. To avoid a proband specific bias and to determine a global expression profile, only those genes were retained which displayed similar responses (up-/down-regulation) in all three probands across all measurements. This analysis resulted in 54 genes (52 induced genes, 2 repressed genes)

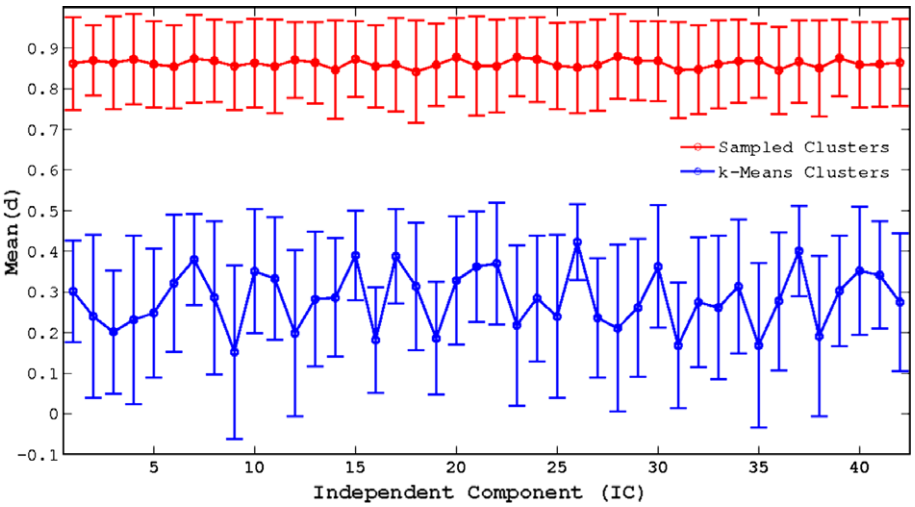

Fig. 1. The means and the standard deviations of the differences d of all clustered row vectors $\mathbf{w} \in \mathcal{W}$ to the corresponding code book vector $\mathbf{c}_p$ for each independent component (IC) compared to a null model of randomly sampled clusters.

indicating that Chemokine signaling, interleukin 1 and TNF-response as well as NFκB signaling are the major pathways strongly influenced by LVS. Prostaglandin synthase 2 and superoxide dismutase 2 are also induced. Lysophospholipase 3 and zinc finger protein 589 are the only repressed genes detected.

3.2. Hierarchical clustering

As a further analysis method, we performed a hierarchical clustering on the data set and selected clusters of differentially expressed genes which show similar time dependent behavior over all three donors. This resulted in 3 clusters corresponding to an early (35 genes), a middle (54 genes) and a late (89 genes) response.

To further define the regulatory network between these genes and to search for interdependent activation waves, Genomatix BiblioSphere analysis was carried out with these data sets. Functional analysis based on the MeSH Filter "Disease" resulted in the following top five terms with good Z-scores for each of the three response terms (Table 1). To gain a focused view on a disease related network, genes related to the top terms of each cluster were combined. This resulted in a network of 49 genes which was analyzed again using BiblioSphere (Fig. 2). The corresponding regulatory network is centered around TNF. As can be seen, the expression levels of genes encoding TNF, as well as TNF-interacting proteins like (TRAF1, TNFAIP8), adhesion molecules (ICAM1) and kinases increase rapidly and decline at later times thus representing an *early response*. At these early times, signal transducer and activator of transcription genes (STAT1/2) are predominantly weakly expressed. In a second signaling wave, the expression levels of TNF induced genes such as the transcription factor NFκB (NFκB1, NFκB2, NFκBIA) and their target genes (IRF7, NUP98, MAPK3K8) increase during an intermediate time interval representing a *middle response*. During a final *late response*, TNF expression declines and expression of the concomitant signaling genes decreases (NFκB1/2, Rel). Late cytokine response, represented by the interferon-induced proteins (IFI2/3, MX1/2), is continually increased during the kinetic experiment. An overlap between these regulatory models and the top 54 genes from the pathway analysis concerning inflammation associated genes like ICAM1, IRAK2, JAG1, NFKB1, NFKB2, TRAF1 and TNF is observed.

Table 1
Terms and Z-scores resulting from a hierarchical clustering and MeSH filtering. ER = early response; MR = middle response; LR = late response. Also the fraction of the genes associated with each MeSH term is given in %.

Resp.	MeSH Term	Z-score	Percentage (%)
ER	Inflammation	53.03	31
ER	Sepsis	24.32	26
ER	Systemic Inflammatory Response Syndrome	22.97	26
ER	Reperfusion Injury	20.86	14
ER	Shock	18.31	20
MR	Inflammation	22.6	9
MR	Cell Transformation, Neoplastic	14.45	17
MR	Cell Transformation, Viral	10.26	7
MR	Leukemia-Lymphoma, T-Cell, Acute, HTLV-I-Assoc.	9.56	2
MR	HTLV-I Infections	8.85	2
LR	Leukemia, Promyelocytic, Acute	155.37	9
LR	Leukemia, Nonlymphocytic, Acute	81.32	12
LR	Leukemia, Myeloid	65.03	15
LR	Leukemia	52.06	18
LR	Translocation, Genetic	42.02	7

3.3. ICA analysis

As a result of the ICA analysis, we obtained $N = M$ expression modes which represent the hypothetical gene regulatory processes. To identify relevant processes represented by the extracted GEMs, we analyzed time dependent patterns formed by the FPs setting up the mixing matrix **A**. To avoid a proband specific bias we filtered out FPs similar among all three probands. Therefore we split up each FP into proband specific temporal patterns and compared them by calculating correlations. Only those FPs which show a high correlation (above 0.8) between all probands specific patterns were used for further analysis. To find FPs comparable to the clusters derived by the hierarchical clustering approach, we identified those with temporal patterns showing high early, middle or late response activity (Fig. 3). We have chosen three FPs for each response type respectively, and merged the extracted gene groups from the corresponding GEMs to three *response groups* (RG) called *early* (149 genes), *middle* (171 genes) and *late* (158 genes).

The biological relevance of these RGs was explored using the Genomatix software. We analyzed each RG using the MeSH Filter "Disease". This resulted in a list of the most related MeSH terms (see Table 2). They are strikingly different to the MeSH terms de-

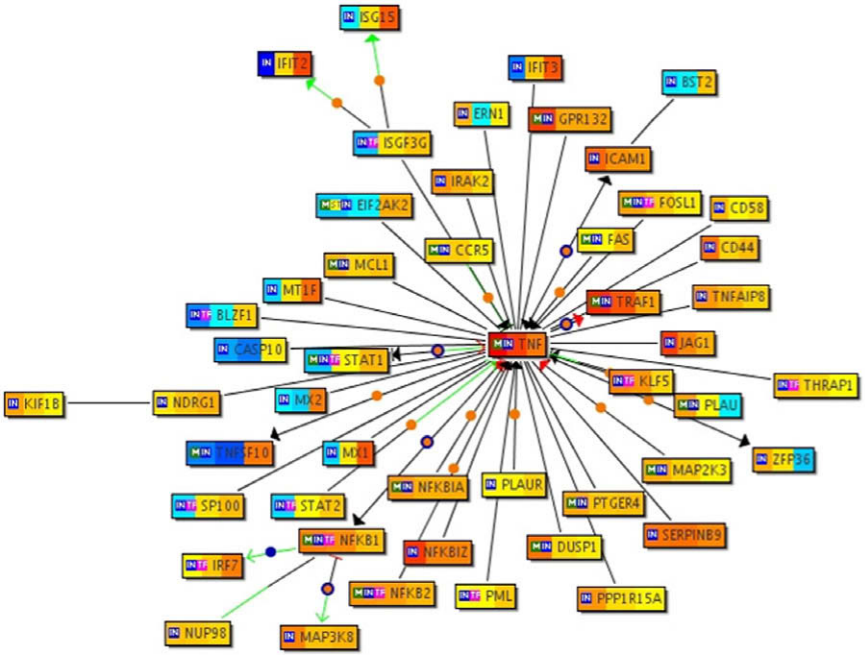

Fig. 2. Functional gene networks resulting from a hierarchical clustering analysis. Expression levels for each gene are color-coded. Overexpression is colored red, underexpression blue. The stripes from left to right code for early, middle and late response. Cited relationships between two genes make up the edges. Display of edges is restricted to those that constitute the shortest path from the central node. If a gene coding for a transcription factor is connected to a gene with a predicted binding site in its promoter, the connecting line is colored green over half of its length near the target gene. Arrowheads at the ends of a connecting line symbolize regulation. Hand-annotated gene–gene relationships are indicated by a circle in the center of the connection line.

rived from hierarchical cluster analysis, and in accordance, the ICA derived terms show noticeably higher *Z*-scores (Inflammation, Systemic Inflammatory Response Syndrome). Furthermore, ICA results show Inflammation as the highest ranked term in all three responses. The percentage of genes associated to MeSH-terms is consistently higher in ICA derived RGs.

The additionally derived network can be seen in Fig. 4. The early response is largely governed by the pro-inflammatory

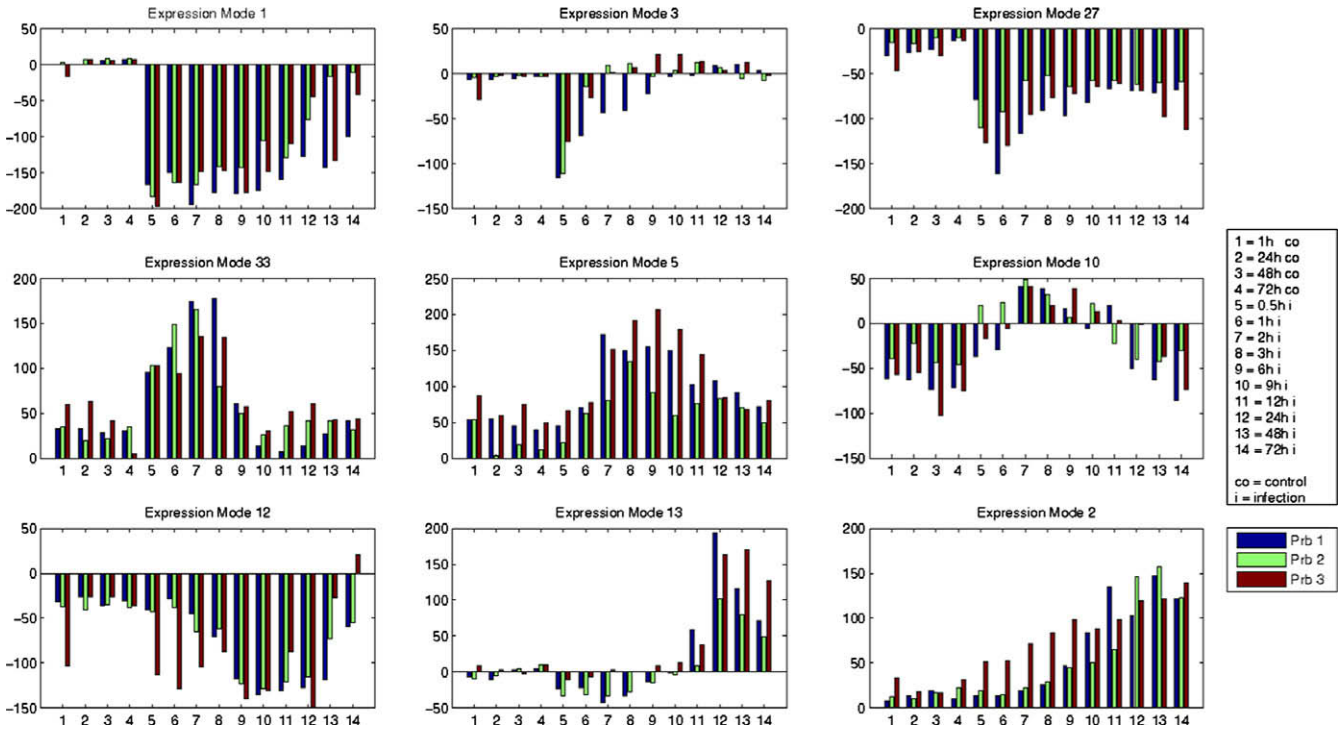

Fig. 3. Feature profiles with similar temporal patterns for all three probands (Prb 1–3). Blue, green and red bars. Shown are only those, used for time dependent response analysis: top: early response, middle: middle response, bottom: late response. Gene *response groups* were created from the corresponding gene expression modes. See text for a detailed explanation.

Table 2
Terms and Z-scores resulting from an ICA analysis and MeSH filtering. ER = early response; MR = middle response; LR = late response. Also the fraction of the genes associated with each MeSH term is given in %.

Response	MeSH Term	Z-score	Percentage (%)
ER	Inflammation	93.74	52
ER	Bacterial Infections and Mycoses	49.36	48
ER	Arthritis	44.51	40
ER	Joint Diseases	43.63	40
ER	Systemic Inflammatory Response Syndrome	42.95	33
MR	Inflammation	64.35	49
MR	Bacterial Infections and Mycoses	30.61	40
MR	Systemic Inflammatory Response Syndrome	27.35	23
MR	Sepsis	25.69	21
MR	Arthritis	24.78	33
LR	Inflammation	46.98	47
LR	Arthritis	27.7	40
LR	Joint Diseases	27.22	41
LR	Rheumatic Diseases	26.15	41
LR	Gram-negative bacterial infections	24.66	30

cytokines (TNF, IL13, IL1B) and chemokines (CXCL2, CXCL3, CXCL5, CCL2-5, CCL8) as well as up-regulation of NFκB. This is followed by activation of TNFα and NFκB induced proteins like TRAF1, MMP9 and the major histocompatibility complex proteins HLA-DRB1, HLA-A and HLA-B. During late response, again the activity of the chemokines CXCL1 and CXCL5 were discovered, as well as the IL8 related genes MRC1, MX1 and CCL18. Here again, the accordance to the 54 top regulated genes is striking through a complete overlap of the associated highest ranked MeSH Terms: "Inflammation", "Arthritis", "Joint Diseases", "Bacterial Infections and Mycoses" and "Systemic Inflammatory Response Syndrome".

A further attribute of ICA based analysis is the grouping of genes into non-exclusive clusters. Hence, genes influencing more than one specific process can be found in more than one RG. Some of those interesting genes are the cytokines IL1B and IL8 or the surface protein coding genes CD36 and CD44 which were identified as presumably key players for gene regulatory networks involved in LVS infection response.

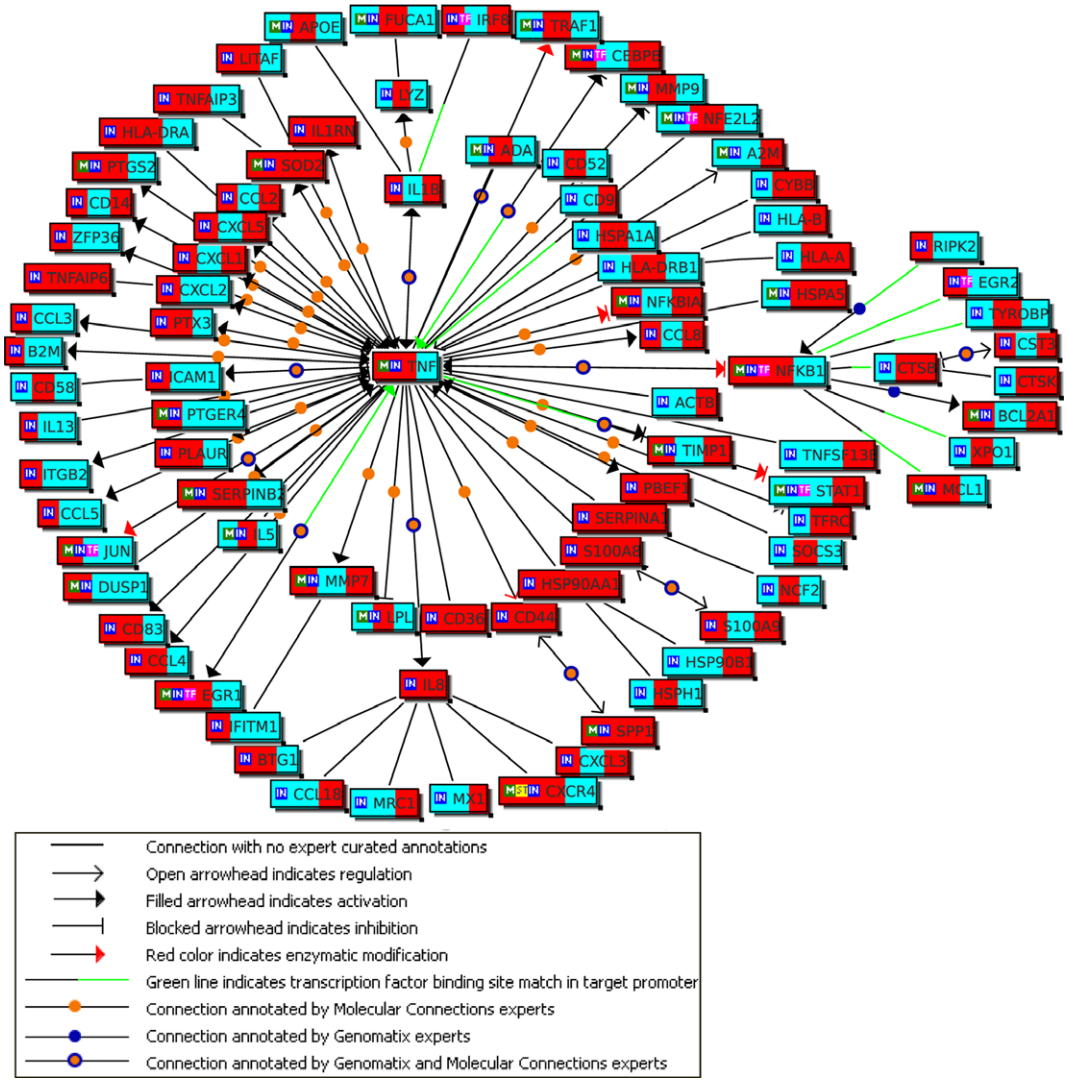

Fig. 4. Functional gene network resulting from the ICA analysis. Stripes from left to right code for early, middle and late response group. If a gene is a member of one or more of the response groups the stripe is colored red. Edges between two genes denote co-occurrence within one abstract. Display of edges is restricted to those that constitute the shortest path from the central node. 'TF' stands for transcription factor, 'ST' means gene is part of Genomatix signal transduction pathway, 'IN' means input gene and 'M' marks a gene which is part of a metabolic pathway.

4. Discussion

Using the data-driven ICA approach, additional novel pathways were identified in addition to pathways similar to the ones deduced from classical hierarchical clustering approaches. Among the early responders, the pro-inflammatory cytokines TNFα and CCL2 were induced, which confirm previous findings about the secretion of large amounts of these inflammatory cytokines in a similar in vitro model using murine macrophages and human cell lines [13]. Furthermore, in a murine macrophage cell line model, testing immediate responder genes by microarray analysis within the first 4 hours after infection with *F. tularensis* LVS, TNFα was found to be the main signal transducer whose expression level was found to be increased along with genes representing cytokine signaling-, enzyme- and transcription factor-families [14]. The differences observed between our early responder genes and the immediate responders found in the murine model system emphasize the need of a multi-time point kinetic model of macrophage response to *F. tularensis* LVS infection with a well established microarray analysis method.

The virulence of *F. tularensis* depends on its ability to escape into the cytosol of the host cell, which reacts with the assembly of the caspase-1 dependent inflammosome complex. This process is closely related to the secretion of IL1b, IL18 and IL33, by which the induction of IL1b was also found with our analysis [15]. Recently, a natural killer (NK) cell cytokine, IFNγ dependent activation pathway was found to be relevant for the specific immune response to *F. tularensis* LVS infection [16]. We found a significant up-regulation of the IFNγ receptor 2 in macrophages, which in turn sensitizes these cells for the NK-cell derived IFNγ to result in a specific response.

These data show that, with the help of in vitro model systems using microarray analysis, the mechanism of *F. tularensis* LVS response can be well characterized and disease specific pathways discovered and identified. Moreover we could show that NFκB plays a major role regulating the immune response to *F. tularensis* LVS infection.

In comparison to the commonly used hierarchical clustering method, we found that our calculations using ICA resulted in higher clustering resolutions. The response specific MeSH terms derived through an ICA analysis are more closely related to the experiment (bacterial infections and mycoses, Gram-negative bacterial infections) and all three *response groups* show Inflammation as the most highly ranked MeSH term. Moreover, the nonexclusive clustering attribute of ICA leads to a more detailed insight into time-dependent patterns of the immune response.

Acknowledgments

The authors gratefully acknowledge Florian Bloechl, Fabian Theis and Philip Wong for proofreading the paper and Dominik Wittmann for useful help on stability analysis. Work was partially supported by a grant of the German Ministry of Defence (M/SAB1/4/A006) and the Helmholtz Alliance on Systems Biology (project CoReNe). Work was performed at the Institute of Clinical Chemistry and Laboratory Medicine in close cooperation with the CIML group, Institute of Biophysics, University of Regensburg.

References

[1] Steke D. Microarray bioinformatics. Cambridge University Press; 2003.
[2] Quackenbush J. Computational analysis of microarray data. Nat Rev Genetics 2001;2:418–27.
[3] Martins RP, Ostermeier GC, Krawetz SW, Draghici S, Kathri P. Global functional profiling of gene expression. Genomics 2003;81:98–104.
[5] Cichocki A, Amari S-I. Adaptive blind signal and image processing. Wiley; 2002.
[7] Hyvärinen A, Oja E. A fast fixed-point algorithm for independent component analysis. Neural Comput 1997;9:1483–92.
[8] Comon P. Independent component analysis, a new concept. Signal Process 1994;36(3):287–314 .
[9] Ostermeier GC, Krawetz SA, Kathri P, Draghici S. Profiling gene expression using onto-express. Genomics 2002;79:266–70.
[10] Langmann T, Schumacher C, Morham SG, Honer C, Heimerl S, Moehle C, et al. ZNF202 is inversely regulated with its target genes ABCA1 and apoE during macrophage differentiation and foam cell formation. J Lipid Res 2003;44(5):968–77.
[12] Friedman N. Inferring cellular networks using probabilistic graphical models. Science 2004;303(5659):799–805.
[13] Loegering DJ, Drake JR, Banas JA, McNealy TL, Mc Arthur DG, Webster LM, et al. *Francisella tularensis* LVS grown in macrophages has reduced ability to stimulate the secretion of inflammatory cytokines by macrophages in vitro. Microb Pathog 2006;41(6):218–25. Epub 2006 Sep 25, 2006.
[14] Andersson H, Hartmanov B, Rydn P, Noppa L, Nslund L, Sjstedt A. A microarray analysis of the murine macrophage response to infection with *Francisella tularensis* LVS. J Med Microbiol 2006;55(Pt 8):1023–33.
[15] Henry T, Monack DM. Activation of the inflammasome upon *Francisella tularensis* infection: interplay of innate immune pathways and virulence factors. Cell Microbiol 2007;9(11):2543–51.
[16] López MC, Duckett NS, Baron SD, Metzger DW. Early activation of NK cells after lung infection with the intracellular bacterium, *Francisella tularensis* LVS. Cell Immunol 2004;232(1-2):75–85.
[17] Gruber P, Theis FJ. Grassmann clustering. Proc. of European Signal Processing Conference; 2006.